学校编码：10384　　分类号____密级_

学号：23220121153057　　UDC_

厦门大学

硕 士 学 位 论 文

真核生物基因组注解及原核生物基因组测序数据研究

Genome Annotation for Eukaryotes and Analysis of Prokaryote Genome Sequencing Data

陈丽娜

指导教师姓名：王颖 副教授

专 业 名 称：系统工程

论文提交日期：

论文答辩时间：

学位授予日期：

答辩委员会主席：_____

评阅人：_____

2015 年 05 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2015 年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2015 年　　月　　日

摘要

生物按照细胞类型分类有真核生物与原核生物，本文主要从真核生物与原核生物两个角度研究生物信息领域的意义所在。

随着高通量测序技术的发展，大量物种被测序并装配获得基因组序列。然而，如何快速准确地注解真核基因组的结构仍然是一个重要问题。目前注解一个真核基因组需要大量来源可靠、不同类型的参考数据源，例如相同或近似物种的蛋白质序列、EST、cDNA 序列以及 RNA-Seq 数据。收集大量可靠的数据，并整合不同数据的分析结果，获得一致、完整的注解结果是一项耗时复杂的工作。因此，本研究第一部分提出一种快速便捷的计算工具 GASS，利用相似物种的注解信息来完成一个新物种基因组的结构注解。首先将相似物种的外显子序列匹配到待注解基因组序列上，将搜寻最合理的转录物注解信息构建为一个动态规划模型，通过最短路径搜索获得最优的注解结果。为了评估 GASS 的性能，我们基于人类基因组注解信息，应用 GASS 注解猕猴基因组， 将注解结果与两个猕猴公共注解数据库(RefSeq、Ensembl)比较，同时利用三个 RNA-Seq 测序数据验证该数据库的精确性。GASS 的注解结果中包含 65%的 RefSeq 外显子与剪切位点。GASS 的敏感性高于 Refseq，与 Ensembl 接近。同时，在基因、外显子、转录物和剪切位点层面，GASS 的特异性高于 Ensembl。本研究同时还发现猕猴 rheMac3 基因组的若干错误拼接位置，这些错误导致了 RefSeq 在外显子边界注解上 2bp 的误差，进而造成了不完整的剪切位点模式。我们通过各种不同的数据源进一步验证了该发现。

此外原核生物群落的多样性是目前研究的热点问题。基于 k-tuple 序列特征的 Alignment-free 方法研究原核生物群落多样性已经得到广泛的应用。然而背景序列建模过程是 k-tuple 特征方法的关键环节。先前基于定阶次马尔科夫模型存在一定缺陷，因此，寻找合适的背景序列模型具有重要意义。因此本文第二部分基于 k-tuple 频度分布设计了 VLMC，IMM，ICM 三种背景序列建模的方法。根

据这三种随机模型，选取不同的微生物群落样本，初步得到了一些结果。实验说明三种模型都有一定的有效性与准确性。

真核生物的基因组注解和原核生物基因组测序数据研究，本文对对这两部分的研究依然存在很多局限性与不足，有待进一步改进。

关键词：k-tuple 频度；随机模型；微生物群落；真核生物；结构注释

Abstract

With the development of high-throughput sequencing techniques, more and more genomes are sequenced and assembled. However, annotating a genome's structure rapidly and expressly remains challenging. Current eukaryotic genome annotations require various, abundant supporting data, such as: species-specific and cross-species protein sequences, ESTs, cDNA and RNA-Seq data. Collecting those data and merging their analytical results to achieve a consistent complete annotation is a complex, time and cost consuming task. In the first part of our study, we propose a fast and easy-to-use computational tool: GASS. It annotates a eukaryotic genome based on only the annotations from another similar species. With the alignments from the exons' sequences of annotated species to the un-annotated genome, GASS detects the optimal transcript annotations with a shortest-path dynamic programming model. For evaluation, GASS is applied to achieve the rhesus annotations. The produced annotations are compared with two existing rhesus annotation databases (RefSeq and Ensembl) directly and with three RNA-Seq datasets. There are more than 65% RefSeq exons and splicing junctions can exactly be found by GASS. The GASS's sensitivity is higher than RefSeq's, and is close to Ensembl's. GASS has higher specificities than Ensembl at gene, transcript, exon and splicing junction levels. We also find the mis-assemblies of rheMac3 genome, which leads to the position shift on exons' boundary and then the incomplete splicing canonical sites in Refseq annotations. They are further approved by various data sources.

Microbial community diversity is the hot topic in ecological studies. Alignment-free methods based on k-tuple frequencies has been widely used. However, how to choose the markov model is the key point for the method. Previously, the k-tuple sequence signature methods based on fixed order MCs. Therefore, it is important to

choose the right model. The second part of this research, we use three models (VLMC, IMM, ICM) to NGS datasets. We get some preliminary results by designing four experiments to analysis the clustering characteristics. The results indicate that all the models are efficacious.

Whether it's the clustering of microbial or eukaryotic genome structure annotation, this studies still have many limitations and shortcomings that needs to be further improved.

Keyword: k-tuple frequency; stochastic model; microbial communities; eukaryotic; genomic structure annotation

目录

摘要.....I

Abstract.....III

第一章 绪论.....1

1.1 研究背景与研究意义.....1

1.2 研究现状综述.....3

1.2.1 真核生物基因组结构注解方法.....3

1.2.2 基于定阶次马尔科夫模型的 k-tuple 频率计算方法.....5

1.3 本文主要工作及创新点.....6

第二章 真核生物基因组注解方法.....8

2.1 基于物种相似性的真核生物基因组注解方法.....8

2.1.1 动态规划算法在基因结构注解中的主要思想.....8

2.1.2 基于动态规划的主要流程.....9

2.2 主要分析流程的代码实现.....14

2.3 实验数据描述.....15

2.4 实验结果分析.....16

2.4.1 实验 1：GASS 实验结果与 RefSeq-rheMac3 直接比较的结果分析.....16

2.4.2 实验 2：基于 RNA-Seq 数据集的结果分析.....22

2.5 探索性发现.....26

2.5.1 实验 3：RefSeq-rheMac3 基因组序列组装错误分析 26

2.6 本章小结 28

第三章 基于 k-tuple 特征原核生物基因组测序数据研究.....30

3.1 k-tuple 思路的主要原理及主要分析流程 30

3.1.1 k-tuple 序列特征方法的核心思想 30

3.1.2 k-tuple 方法的分析流程 31

3.2 VLMC，IMM，ICM 三种模型描述 37

3.2.1 VLMC 模型描述 37

3.2.2 IMM 模型描述 39

3.2.3 ICM 模型描述 40

3.3 实验数据描述 42

3.3.1 细菌基因组数据描述.....42

3.3.2 海洋微生物（硅藻）转录组数据描述.....44

3.3.3 海洋微生物宏转录组与宏基因组数据描述.....45

3.3.4 全球海洋宏转录组样本数据描述.....46

3.4 实验结果与分析 48

3.4.1 基于细菌基因组样本的聚类分析.....48

3.4.2 基于海洋微生物宏转录组样本聚类分析.....49

3.4.3 基于海洋微生物宏转录组与宏基因组样本聚类分析.....52

3.4.4 基于全球海洋宏转录组样本聚类分析.....53

3.5 本章小结 55

第四章 总结与展望 55

【参考文献】.....58

在学期间发表以及完成的论文 65

致谢语.....66

Contents

Abstract in Chinese ..... **I**

Abstract in English ..... **III**

Chapter 1 Introduction **1**

1.1 Research background and significance **1**

1.2 The Review of study status **3**

1.2.1 Genome annotation methods for eukaryotes 3

1.2.2 The transition probabilities based on fixed order markov model 5

1.3 Main and innovation works in this thesis **6**

Chapter 2 Genome annotation methods for eukaryotes **8**

2.1 Genome Annotation for Eukaryotes Based on Species Similarity **8**

2.1.1 The main ideas of dynamic programming model 8

2.1.2 The main process of dynamic programming model 9

2.2 Implementation of the main process **14**

2.3 Descriptions of datasets **15**

2.4 Expermental results and analysis **16**

2.4.1 Experiment 1: Direct comparison with RefSeq-rheMac3 16

2.4.2 Experiment 2：Comparison based on RNA-Seq datasets 22

2.5 Further investigation of RefSeq-rheMac3 ..... **26**

2.5.1 Experiment 3：Mis-assembly of rheMac3 genome 26

2.6 Brief summary **28**

Chapter 3 Analysis prokaryote genome sequencing data based on k-tuple frequency 30

3.1 Methods based on k-tuple frequency 30

3.1.1 The main theoretical of k-tuple frequency.....30

3.1.2 The main process of k-tuple frequency.....31

3.2 Models based on VLMC，IMM，ICM.....37

3.2.1 Descriptions of VLMC model 37

3.2.2 Descriptions of IMM model.....39

3.2.3 Descriptions of ICM model 40

3.3 Experiments of metatranscriptomic data 42

3.3.1 Descriptions of bacterial samples 42

3.3.2 Descriptions of metatranscriptomic samples 44

3.3.3 Descriptions of metagenomic and metatranscriptomic samples.....

.....45

3.3.4 Descriptions of global ocean metatranscriptomic samples.....46

3.4 Expermental results and analysis 48

3.4.1 The results based on bacterial samples 48

3.4.2 The results based on metatranscriptomic samples 49

3.4.3 The results based on metagenomic and metatranscriptomic samples.....52

3.3.4 Descriptions of global ocean metatranscriptomic samples.....53

3.5 Brief summary.....55

Chapter 4 Conclusions and Future Works 55

【References】 58

Published and finished papers during the study period 65

Acknowledgements 66